

CHEMOMETRICS AS AN AID IN AUTHENTICATION

Ramón Aparicio

Instituto de la Grasa, Avda. P. Garcia Tejero, 4 41012 Sevilla – Spain

Tel: 00-34-95-4611550 Fax: 00-34-95-4616790 e.mail: aparicio@cica.es

CONTENTS

9.1. Introduction

9.2. Chemometric procedures in food authentication

9.2.1. Pre-treatment of data

9.3. Multivariate procedures

9.3.1. Cluster Analysis

9.3.2. Factor Analysis

9.3.3. Multidimensional Scaling

9.3.4. Discriminant Analysis

9.3.5. Regression analysis

9.4. Artificial Intelligence methods in food authentication

9.4.1. Expert Systems

9.4.2. Neural Networks

9.4.3. Fuzzy Logic

9.1. Introduction

Thousands of chemical compounds have been identified in oils and fats although only a few hundreds are used in authentication. This means that each object (food sample) may have a unique position in an abstract n-dimensional hyperspace. A concept that is difficult to interpret by analysts as a data matrix exceeding three or more features means already a problem. The art of extracting chemically relevant information from data produced in chemical experiments, by means of statistical and mathematical tools, is given the name of Chemometrics. It is hence an indirect approach to the study of the effects of multivariate factors (or variables) and hidden patterns in complex data sets. Chemometrics is habitually used for (a) exploring patterns of association in data, and (b) preparing and using multivariate classification models. The arrival of Chemometrics techniques has allowed the quantitative as well as qualitative analysis of multivariate data and, in consequence, it has allowed that we can now analyse and model a lot of different types of experiments.

The general concept of authentication, on the other hand, is not only circumscribed to adulteration but also include aspects as characterisation, geographical origin of foodstuffs, extraction and processing systems among others. Thus, the combined effect of a great amount of data, supplied by the current sophisticated instrumentation, and the various issues included inside the concept of authenticity means that the analyst should face a complex work. Thus, we cannot think that we only need to put data in one side of a 'black box' (the computer) and get the results from the other side.

9.2. Chemometric procedures in food authentication

The chemometric procedures that are currently applied in empirical investigations have been notably improved in the recent years with the assistance of the computer science. Researchers have passed from the initial application of univariate analyses to the extensive use of multivariate procedures in less than one decade. This qualitative step has been possible because the new sophisticated analytical instruments are now able to analyse dozens of chemical compounds in hundreds of samples daily, and secondly due to the personal computers that can work with a great diversity of software packages.

The application of mathematical algorithms has allowed conclusions to be arrived at that were unthinkable only a few decades ago. However, the conclusions will be useful only if the analysts follow strictly the next three steps:

a) *Exploratory Data Analysis* (EDA). This analysis is also called “pre-treatment of data” and is essential to avoid wrong or obvious conclusions. The EDA objective is to obtain the maximum useful information from each piece of chemical/physical data because the perception and experience of a researcher cannot be sufficient to single out all the significant information. This step comprises descriptive univariate statistical algorithms (e.g., mean, normality assumption, skewness, kurtosis, variance, coefficient of variation), detection of outliers, cleansing of data matrix, measures of the analytical method quality (e.g., precision, sensibility, robustness, uncertainty, and traceability) (Eurachem 1998), and the use of basic algorithms such as Box-and-Whisker, Stem-and-Leaf, etc.

b) *Bivariate statistics*. The objective here is to look for possible relationships between pairs of variables. Pearson’s correlation has traditionally been the most used although the analysis of the correlation matrix should be studied before the use of most multivariate statistical procedures.

c) *Multivariate statistics*. The main matter facing researchers in the authentication of foods characterised by numerous variables is how to organise the observed data into meaningful structures. Independent of the names given to the multivariate procedures, they can be clustered into two basic groups. The descriptive (non-supervised) group is characterised by the fact that there is not any previous hypothesis classifying or defining the objects (samples). The procedures clustered inside this group analyse the information given and explore the data matrix in the searching of new knowledge. There are many statistical procedures inside this group, factorial analysis (e.g. principal component analysis and maximum likelihood), canonical correlation, multiple partial correlation, clusters, correspondence analysis and multidimensional scaling among other less applied. The explanatory (supervised) procedures, on the contrary, have the objective of checking *a priori* hypothesis or they are simply dependence models that can

be subsumed under the general concept of regression. Discriminant analyses, multivariate analysis of variance, log-lineal models are, among many others, explanatory procedures also.

9.2.1. Pre-treatment of data

Chemometrics and Artificial Intelligence procedures require of a strict process of data selection. Since almost all the statistical procedures are based on Bayesian theory, the information has to comply with eight conditions:

- a) The error of the method (or overall variability) and uncertainty should be known.
- b) Chemical and/or physical analyses should be carried out in triplicate, or at least duplicate.
- c) The assumption of normality should be verified with each variable.
- d) Outliers should be detected and corrected, if possible.
- e) The data set should have information of the whole range of values of the knowledge domain to be studied (e.g., the cultivars of a vegetable oil).
- f) The validation (test) set should be independent of training set, if possible.
- g) The standardisation of data should agree with the objective to be studied.
- h) The number of variables (e.g., chemical compounds) should be lower than objects (e.g., samples).

Knowing which factors contribute to the overall variability it would be possible to improve the analytical methodology. The whole error (first condition) is composed of the systematic error, or bias, the unspecified random error, and a series of errors produced during the chemical or physical analyses. Uncertainty, also expressed as standard deviation (type A of uncertainty), is the concept for measuring the quality of the analytical procedures (Taylor and Kuyaat 1994). To implement the statistical procedures the analyst need parameters quantified in triplicate, or at least in duplicate.

A great majority of statistical procedures are based on the assumption of normality of variables, and it is well known that the central limit theorem protects against failures of normality of the univariate algorithms. Univariate normality does not guarantee multivariate normality though the latter is increased if all the variables have normal distributions and, in any case, it avoids the deleterious consequences of skewness and outliers on the robustness of many statistical procedures. Anyway, numerous transformations are able to reduce skewness or the influence of outlying objects.

Due to Chemometrics works almost exclusively with numbers, abnormal data can lead the mathematical procedures to obvious or wrong conclusions; for example, outliers inside the training set of neural networks. Most of the outliers can be detected and some of them corrected before applying the definite mathematical procedures by the so-called robust algorithms (Armstrong and Beck 1990) although most statisticians usually remove them when the database is large enough. Those wrong conclusions are, however, mostly due to databases that do not keep all aspects of the food characterisation (e.g. partial or skew databases) or do merge data of chemical compounds quantified with different techniques (e.g. fatty acids quantified by different chromatographic columns). A classical example is an authentication of the geographical origin of Italian virgin olive oil samples by ANN (Zupan *et al.* 1994). The differences between oils were mainly due to the use of different chromatographic columns (packed columns and open tubular columns vs. capillary columns) when quantified FFA of the oils from Southern and Northern Italy respectively. The neural network has therefore thus mostly learned to recognise the chromatographic columns.

Databases should be split into two independent parts: (a) the training set and (b) the test set. The first set is used to obtain the mathematical equations while the second set is used to validate the equations. The selection of objects (samples) for these sets should be carried out with random methods (e.g., random numbers) and the independent test set should have at least 25% of total samples, and the samples of both sets must contain exemplars encompassing the appropriate variance over all relevant properties for the problem at hand. The use of an external validation set does not invalidate the use of any internal validation (e.g., cross-validation or leverage correction) that has been applied with success there where the total number of objects was small. At this stage it is of great importance to have the correct exemplars in the training and test sets because if outliers

are included in the construction of a model then inaccuracies in the predictions from new multivariate data are likely to occur.

Chemical or physical data can differ by orders of magnitude (e.g., ppb or ppm) and, moreover, data are collected from instruments that often give information in different scientific measurements (e.g., °C, % or g/l). In these cases, the scaling (also called standardisation) should be applied in order to re-adjust the individual contributions to the outcome on an equal basis, so avoiding that some variables can weight more than other on the results. Furthermore, data standardisation make residuals more symmetrically distributed that is important because least square estimation (LS) is consistent with non-random residuals.

Overfitting is the commonest problem in multivariate statistical procedures when the number of variables is greater than objects (samples); "one can fit an elephant" with enough variables. Tabachnick and Fidell (1983) have suggested the minimum requirements for some multivariate procedures, otherwise overfitting or underfitting can occur in a somewhat unpredictable manner regardless of the multivariate procedure chosen.

9.3. Multivariate procedures

The main goal of this section is to provide a summary of several of the most widely used multivariate procedures in food authentication out of the vast array currently available. They are included inside well-known computer packages as BMDP, IMSL, MATLAB, NAG, SAS, SPSS and STATISTICA. The first three subsections describe unsupervised procedures, called exploratory data analysis also, that can reveal hidden patterns in complex data by reducing data to more interpretable information. That information emphasizes the natural grouping in the data and show which variables most strongly influence those patterns. The four and five subsections are focused on the supervised procedures of discriminant analysis and regression. The former produces good information when applies under the strictness of certain tests whereas the latter is mainly used when the objective is the calibration.

9.3.1. Cluster Analysis

The term cluster analysis comprises classification algorithms designed to understand the information of data matrices, to describe the similarities and dissimilarities among objects (samples) and to single out categories grouping similar objects (Hartigan 1975; Zupan 1982). This collection encompasses the following algorithms: k-means clustering, that minimises within-cluster variability while simultaneously maximises between-cluster variability (Jacobsen and Gunderson 1986), block clustering, that simultaneously amalgamates objects and variables, and tree clustering called joining clustering as well. The latter algorithm, that is the most popular (Massart and Kaufman 1983), is based on two kinds of sub-procedures: distance measures and amalgamation rules.

The hierarchical clustering method uses the distances (or dissimilarities) between variables when forming the clusters. The distances that can be computed are based on a single dimension or multiple dimensions:

- a) The Euclidean distance is the geometric distance in the multidimensional space, and it is probably the most commonly chosen,
- b) The Squared Euclidean distance is similar to the preceding one though it adds progressively greater weight on objects that are further apart.
- c) The Chebychev's distance is suggested to state that there are objects as "different" on any one of the dimensions.
- d) The Power distance is applied when the analyst wants to increase, or decrease, the progressive weight on each dimension.
- e) The Manhattan (city-block) distance is simply the average difference across dimensions. If this distance is applied, then the outliers should be previously removed or corrected because the effect of outliers is dampened with this distance.
- f) The Percent disagreement is a distance particularly useful if the information is categorical in nature.

g) The Mahalanobis' distance is a measure between two points in the space defined by two or more correlated variables. If the variables are correlated, the Mahalanobis distance will adequately account for the correlation whereas the simple Euclidean distance is not an appropriate measure. The results of Mahalanobis' distance will be identical to the Euclidean distance if the variables are uncorrelated.

Once several objects have been linked together, the next step is to determine the distances between those new clusters. This new procedure is carried out by linkage or amalgamation rules that determine when two clusters are similar enough as to be linked together. There are various possibilities:

a) The K-nearest neighbour (single linkage) is determined by the distance of the two closest objects in the different clusters. Thus, the resulting clusters tend to represent long "chains".

b) The K-furthest neighbour (complete linkage) is determined by the greatest distance between any two objects in the different clusters. This amalgamation method performs quite well with naturally distinct objects (or variables) but is inappropriate if the clusters tend to be elongated.

c) The unweighted pair-group average is the average distance between all pairs of objects in two different clusters. This amalgamation method is not affected by the shape of clusters and, hence, it should be used when the objects form natural distinct groups.

d) The weighted pair-group average should be used when the cluster sizes are suspected to be largely uneven. It is identical to the unweighted pair-group average method, except that the size of the respective clusters is used as a weight.

e) The unweighted pair-group centroid it is the centre of gravity for each cluster and the distance between two clusters is determined as the difference between centroids.

f) The weighted pair-group centroid is appropriate when there are appreciable differences in cluster sizes. It is identical to the previous one, except that it takes into consideration differences in cluster sizes.

g) The Ward's method is different from all other methods because it uses the approach of the analysis of variance to evaluate the distances between clusters. This method is regarded as very efficient although it tends to create clusters of small size.

Clustering analysis has been applied to many authentication problems. Thus, the authentication of European monovarietal virgin olive oils have been analysed using various chemical compounds of the unsaponifiable matter (Aparicio and Alonso 1994) and volatile compounds (Aparicio *et al.* 1997; 2000). The volatiles used for the example (Figure 1a-1b) have not been explicitly selected to distinguish monovarietal oils but to point out possible differences between the chemometric procedures with respect to their results. The figures show how the election of the amalgamation rules and the linkage distances determine the result, so pointing out that analysts should meditate this subject before applying clustering analysis.

9.3.2. Factor Analysis

The techniques for the modelling of complex data are clustered inside the factor analysis (FA), principal components analysis (PCA) being the most applied in authentication. The objective of applying FA is to obtain a number of unobservable factors, from the original set of observable variables (e.g., chemical components or wavenumbers), so as to reduce a large raw data matrix to one smaller while retaining most of the original information. FA produces several linear combinations of observed variables generally called eigenvectors. The process of FA includes selecting a group of original variables, building the correlation matrix, determining the number of eigenvectors to be considered, extracting a set of eigenvectors from the correlation matrix, rotating the eigenvectors to increase interpretability and, eventually, making the conclusions. Conclusions should take into consideration that: (i) they must “make sense”; (ii) they should be supported, when used in authentication, by the understanding from another science (Chemistry, Physics, Biochemistry, Agronomy, etc) so demonstrating that results were not attained by chance.

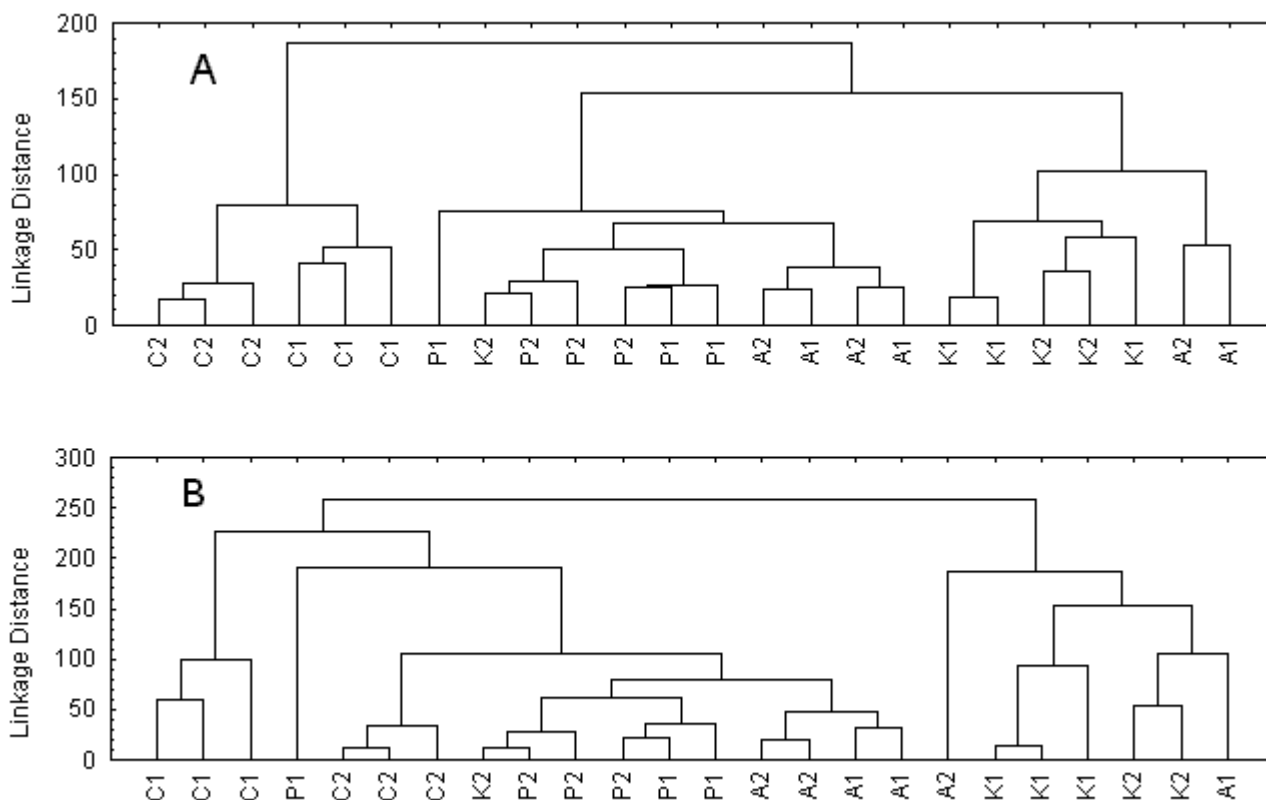


Figure 9.1. Authentication of monovarietal virgin olive oils: Results of applying Clustering Analysis to volatile compounds. The Mahattan (city-block) distance metric and Ward's amalgamation method were used in (A), and Squared Euclidean distance and complete linkage amalgamation method were used in (B).

Note: A, cv. Arbequina (6); C, cv. Coratina (6); K, cv. Koroneiki (6); P, cv. Picual (6); 1, harvest 1991; 2, harvest 1992. Olives were harvested at three levels of maturity (unripe, normal, overripe).

(Source: SEXIA™ Group – Instituto de la Grasa Seville Spain)

The first step of FA is the factor extraction; those described below are the most common. The extraction methods calculate a set of orthogonal factors (or components) that in combination reproduce the matrix of correlation. The criteria used to generate the factors are not homogeneous for all methods but the differences between their solutions may be quite small.

- a) Principal components analysis (PCA) is the most used multivariate procedure as it is easy to interpret and permits an explanation for the maximum variability of initial distribution. Moreover, there is no need to invert a matrix when applying PCA.

- b) The analysis of communalities by multiple R-square is used for estimating the communalities. Prior to factoring, the diagonal of the correlation matrix (communalities) will be computed as the multiple R-square of the respective variable with all other variables.
- c) Iterated communalities method uses multiple R-square estimates for the communalities. The method adjusts the loadings over several iterations using the residual sums of squares to evaluate the goodness-of-fit of the resulting solution.
- d) Centroid method is the geometrical approach to FA (Wherry 1984).
- e) Principal axis method firstly computes the eigenvalues from current communalities, and then the next communalities (sum of squared loadings) are repeatedly recomputed, based on the subsequent extracted eigenvalues and eigenvectors, to get minimum changes in communalities.
- f) Maximum likelihood factors calculate the loadings and communalities that maximise the likelihood of the correlation matrix. It needs an *a priori* hypothesis on the number of possible factors.

Although each one of those methods of extracting factors has a distinct mathematical background, the analyst should use an algorithm that removes those eigenvectors that are dependent enough on the actual objects because this fact could reduce the validity of the model. Cross-validation (Wold 1978) is the most used for selecting eigenvectors that give the minimum residual sum of squares for the omitted objects (Piggott and Sharman 1986). An additional check for the appropriateness of the extracted factors is to check deviations between reproduced and observed matrices, the new matrix being the matrix of residual correlations. This matrix may point to particular correlation coefficients that cannot be reproduced appropriately by the current number of factors.

The graphical representation is also important to visualize the results attained by the methods that extracted the factors. The objective is not to improve the fit between the observed and reproduced correlation matrices but to have an aid to the interpretation of scientific results, making them more understandable. Eigenvector rotation is the most used, and the four types of rotation are as follows:

- a) Varimax normalised method that is the most commonly used, performs a rotation of the normalised factor loadings; raw factor loadings divided by the square roots of the respective communalities. This rotation maximises the variances in the columns of the matrix of the squared normalised factor loadings.
- b) Quartimax normalised method maximises the variances in the rows of the matrix of the squared normalised factor loadings, whereas if the normalisation has not been requested, Quartimax raw, the maximisation is done in the squared raw factor loadings.
- c) Equamax rotation is a weighted mixture of Varimax and Quartimax rotations. It simultaneously maximises the variances in the rows and columns of the matrix of the squared raw factor loadings.
- d) Oblique rotation has been developed to rotate factors, without the constraint of orthogonality of factors although they are often not easily interpreted (Wherry 1984).

From an empirical point of view, three practical issues must be considered with the application of FA: (1) multicollinearity and singularity, (2) outliers among variables and with respect to the solution, and (3) validation of results. Extreme multicollinearity and singularity must be avoided for those algorithms that require matrix inversion. When multicollinearity is present (Tabachnick and Fidell 1983) then it may be necessary to eliminate some variables. Variables that are unrelated to others should be identified as potential outliers. To determine which objects (samples) are multivariate outliers, one should calculate a critical value by looking up critical χ

- d) Oblique rotation has been developed to rotate factors, without the constraint of orthogonality of factors although they are often not easily interpreted (Wherry 1984).

From an empirical point of view, three practical issues must be considered with the application of FA: (1) multicollinearity and singularity, (2) outliers among variables and with respect to the solution, and (3) validation of results. Extreme multicollinearity and singularity must be avoided for those algorithms that require matrix inversion. When multicollinearity is present (Tabachnick and Fidell 1983) then it may be necessary to eliminate some variables. Variables that are unrelated to others should be identified as potential outliers. To determine which objects (samples) are multivariate outliers, one should calculate a critical value by looking up critical χ^2 at the desired α -level (Tabachnick and Fidell 1983). Confirmatory FA is performed to test hypothesis about the structure of underlying processes (e.g., which variables in a data set form coherent clusters that are relatively independent of one another). Thus, confirmatory FA needs a validation process that can be carried out with an independent test data (external validation) or the same data set (validation set). The algorithms used in the latter validation include cross-validation, leverage correction, bootstrap or Mallows C_p (Martens and Naes 1989; Tabachnick and Fidell 1983).

On the other hand, if the correlation matrix has variables that are 100% redundant, then the inverse of the matrix cannot be computed; it is the so-called ill-conditioning matrix. This happens when there are high-intercorrelated variables (e.g., a variable that is the sum of two other variables). The statistical packages can artificially add a small constant to the diagonal of the matrix, thus lowering the whole correlation in the correlation matrix, but analysts usually forget the resulting estimates will not be exact.

In practice, there are not great differences among the methods for extracting factors. Thus, PCA could be suggested in authentication studies because it is simply a mathematical transformation of the raw data. Figure 9.2 illustrates the case of monovarietal virgin olive oils characterised by volatile compounds described in the cluster paragraph.

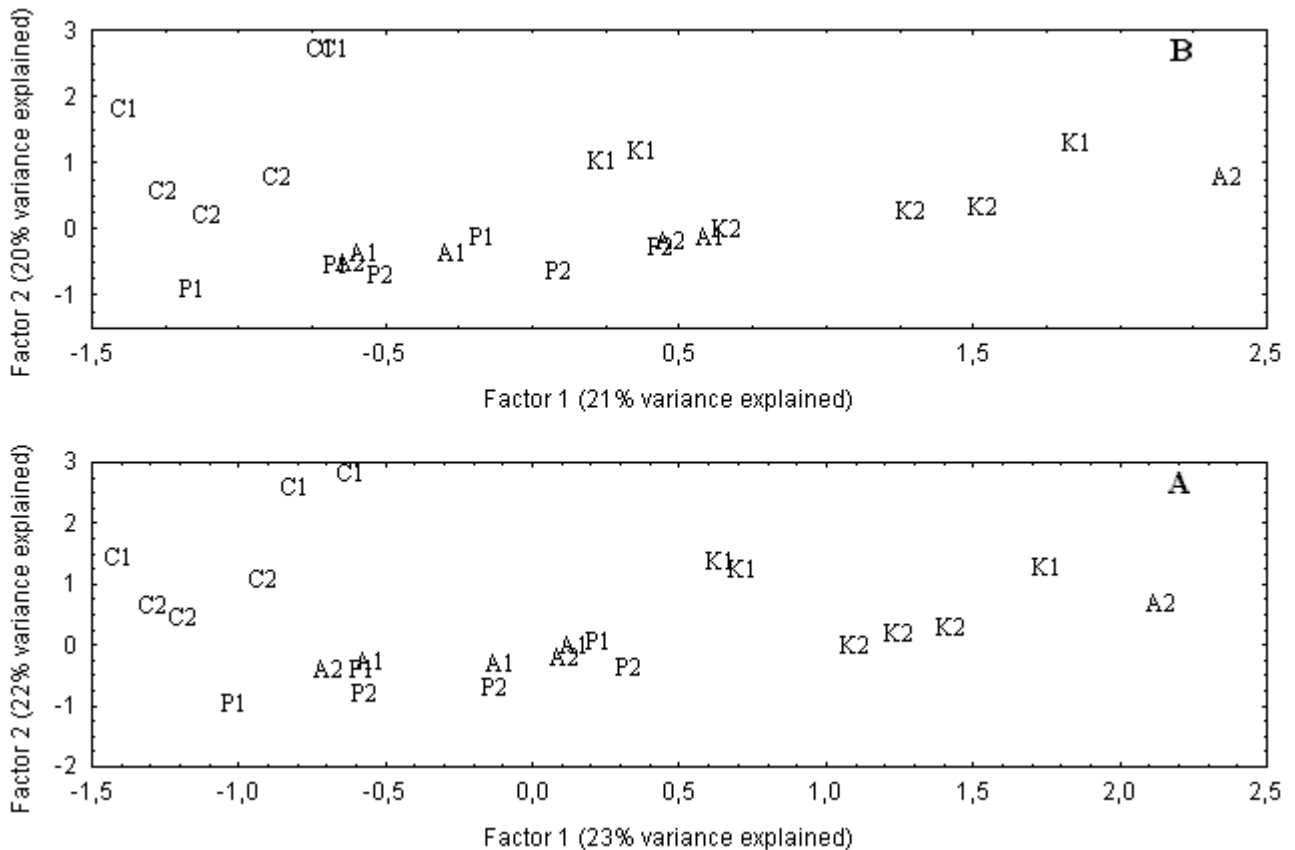


Figure 9.2. Authentication of monovarietal virgin olive oils: Results of applying Factor Analysis to the volatile compounds. (A) Maximum Likelihood and Varimax rotation. (B) Principal Components and Varimax rotation. *Note:* A, cv. Arbequina; C, cv. Coratina; K, cv. Koroneiki; P, cv. Picual.

(Source: SEXIA™ Group – Instituto de la Grasa Seville Spain)

9.3.3. Multidimensional Scaling

Multidimensional scaling (MDS) is an alternative to factor analysis when the goal of the analysis is to authenticate foodstuffs based on observed distances (similarities and dissimilarities) between investigated objects (Borg and Lingoes 1987), in addition to correlation matrices (Shiffman *et al.* 1981). MDS is not so much an exact procedure as rather a way to "rearrange" objects in an efficient manner. The program uses a minimisation algorithm that evaluates different configurations with the goal of maximising the goodness-of-fit (or minimising "lack of fit"). The stress measure is used to evaluate how well (or poorly) a particular configuration reproduces the observed distance matrix. Thus, the smaller the stress value, the better is the fit of the reproduced distance matrix to the observed distance matrix. An interesting application is given in Aparicio *et al.* (1996a).

The analyst should check the Shepard diagram that represents a step line so-called D-hat values. If all reproduced distances fall onto the step-line, then the rank ordering of distances (or similarities) would be perfectly reproduced by the dimensional model while deviations from the step-line mean lack of fit. The interpretation of the dimensions usually represents the final step of this multivariate procedure. As in factor analysis, the final orientation of axes in the plane (or space) is mostly the result of a subjective decision by the researcher since the distances between objects remain invariable regardless of the type of the rotation. However, MDS and FA are different methods. FA requires that the underlying data be distributed as multivariate normal whereas MDS does not impose such restriction. MDS often yields more interpretable solutions than FA because the latter tends to extract more factors. MDS can be applied to any kind of distances or similarities (those described in cluster analysis), whereas FA requires firstly the computation of the correlation matrix. Figure 9.3 shows the results of applying MDS to the samples described in CA and FA paragraphs.

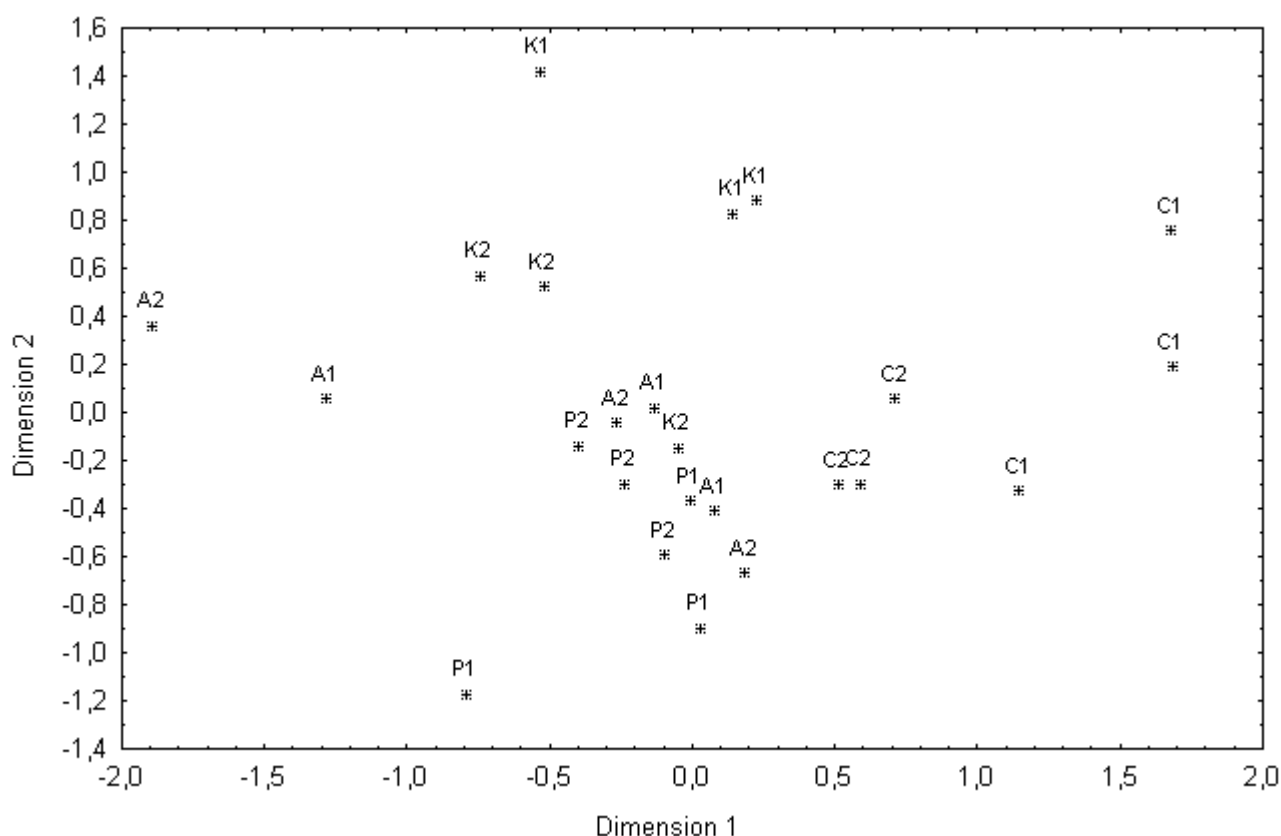


Figure 9.3. Authentication of monovarietal virgin olive oils: Results of applying Multidimensional Scaling to volatile compounds. The distance was Manhattan (city-block) and the amalgamation was Ward's method.
Note: A, cv. Arbequina; C, cv. Coratina; K, cv. Koroneiki; P, cv. Picual.
 (Source: SEXIA™ Group – Instituto de la Grasa Seville Spain)

9.3.4. Discriminant Analysis

Analyses of the above type belong to the category of “unsupervised learning” whereas discriminant analysis falls into the “supervised” analyses of multivariate data. Thus discriminant analysis is used not only to determine which variables discriminate between naturally occurring groups but also which variables are the best predictors discriminating between groups. The procedure can be interpreted as a special type of factor analysis extracting orthogonal factors (Cooley and Lohnes 1971) although it is disparaged by certain chemometricians who think that it capitalises on chance because it “picks and chooses” the variables to be included in the model. The procedure can render results as good as the unsupervised procedures if it is applied with rigour: Fisher for the selection of initial variables, adequate values of F-to-Enter and F-to-Remove, Jackknife algorithm (internal validation test) and an external validation set. But before applying the procedure, the analyst should inspect the means and standard deviations or variances of each group to detect outliers. They should be removed or methods used to correct their influence; a few extreme outliers have a large impact on the means and increase the variability as well. Another assumption of discriminant function analysis is that the variables to be used discriminating between groups cannot be completely redundant. If any one of the variables is completely redundant with respect to the other variables then the matrix is “ill-conditioned” and cannot be inverted.

The main objective of this procedure is to build a model in which the selected variables can predict to which group an object (sample) belongs; for example, a sample of Coratina virgin olive oil with respect to the Coratina cultivar group by a selected set of volatile compounds. The selection of variables to build the model should be done step-by-step. At each step, all unselected variables are evaluated to know which one contributes most to the discrimination between groups, and that variable is then included in the model. The backward stepwise analysis first includes all variables in the model and then, at each step, eliminates the variable that contributes least to the prediction of group membership. The model keeps only the important variables, that is, those variables that contribute the most to the discrimination between groups. The control of the variables included or excluded from the model is carried out by the *a priori* F-to-enter and F-to-remove values. These F-values are a measure of the extent to which a variable makes a unique contribution to the prediction of group membership, and must be selected with strictness;

the higher the values, the lesser the variables included but the better the validation results. Thus, those F-values are taken from an F-distribution table ($m \times n$) at 99%, where “m” is the number of groups and “n” is the number of samples of the smallest group.

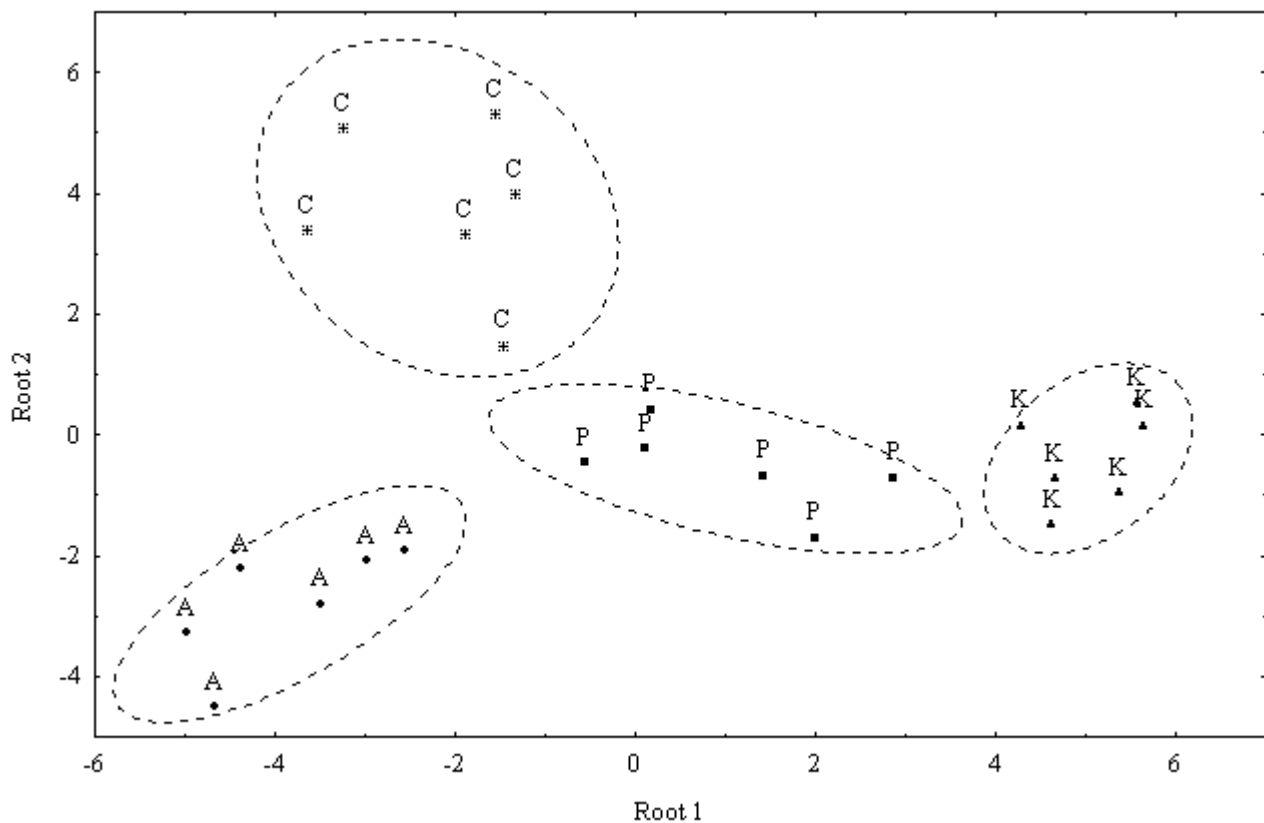


Figure 9.4. Authentication of monovarietal virgin olive oils: Results of applying Stepwise Linear Discriminant Analysis to volatile compounds. Classification was carried out by four volatiles: (E)-2-hexenal, butyl acetate, (E)-3-hexenal, 2-methyl-3-buten-2-ol. F-to-Enter was 8.0; Tolerance was upper 0.52 for all selected volatiles. *Note:* A, cv. Arbequina; C, cv. Coratina; K, cv. Koroneiki; P, cv. Picual.

(Source: SEXIA™ Group – Instituto de la Grasa Seville Spain)

Before assuming the final conclusions, it is probably a good idea to review the within-groups variances and correlation matrices, and re-run the analyses excluding those groups that are of less interest when in doubt. The *a priori* probabilities are an additional factor that needs to be considered when analysing the objects to be authenticated; if there are more objects (samples) in one group than in any other, the discriminant analysis should be readjusted with *a priori* probabilities proportional to the sizes of the groups. Once the procedure has computed the classification scores, the analyst should know the posterior probabilities by the Mahalanobis' distance. The probability that an object can be classified into a particular group is proportional to the Mahalanobis' distance from the sample location to the group centroid.

Figure 9.4 shows the results of applying SLDA to the same samples used in the previous chemometric procedures.

A good alternative to the use of supervise procedures to cluster samples, as for instance SLDA does, is to select the best variables characterising *a priori* groups by SLDA and then to run PCA exclusively with those selected variables (Baeten *et al.* 1996). This two-step procedure would avoid building a model from pure noise or where noise had a great influence as PCA could do under certain circumstances. Figure 9.5 shows the result of applying PCA on the four volatile compounds selected by SLDA. The groups of monovarietal virgin olive oils are much more clear after the process described than applying PCA to all volatile compounds. This means that the first two principal components of the latter procedure contained noise and hence the group classifications were worse.

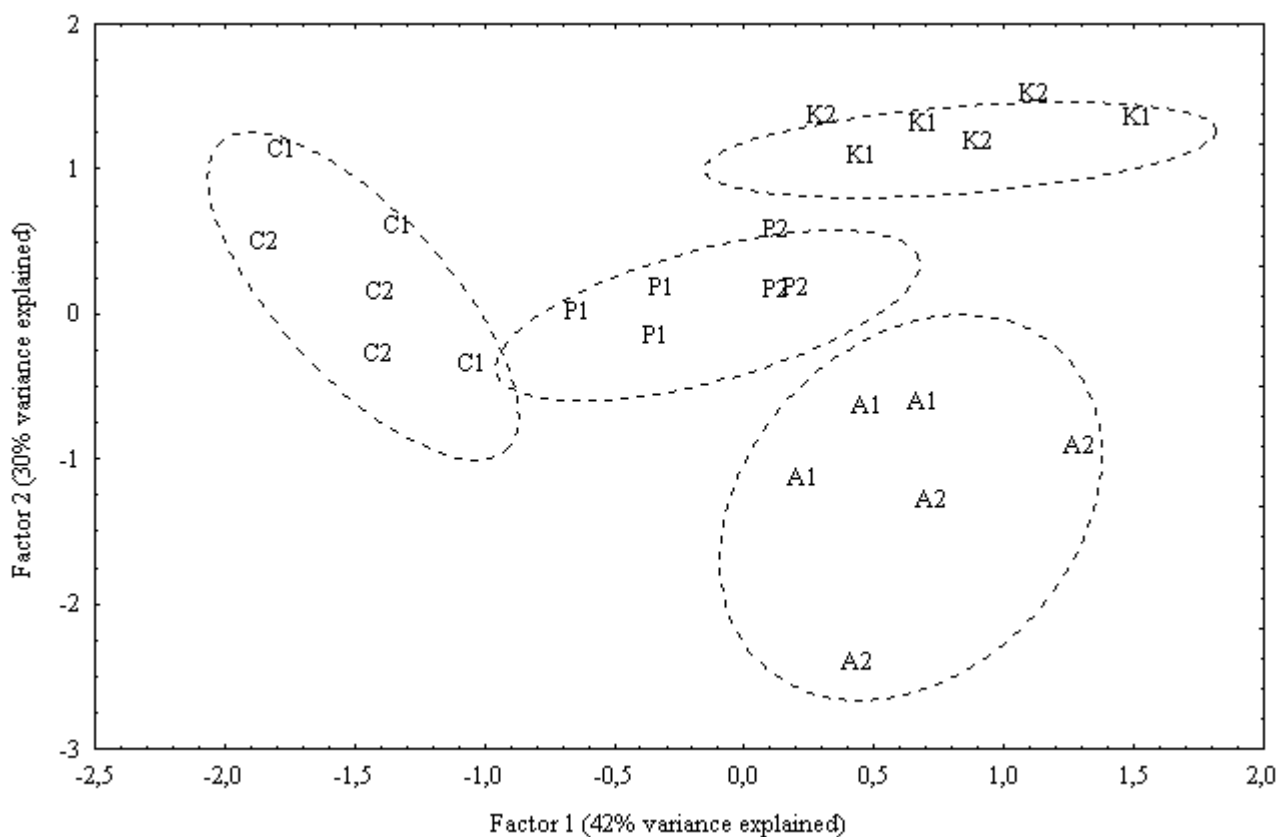


Figure 9.5 Principal Components Analysis applied to volatile compounds selected by Stepwise Linear Discriminant Analysis.

(Source: SEXIA™ Group – Instituto de la Grasa Seville Spain)

9.3.5. Regression procedures

In many applications, it is expensive, time consuming or difficult to measure a property of interest directly, and analyst should predict it based on other properties that are easier to measure. The objective is to design a model for the relationship. From the mathematical point of view, multivariate calibration is developed for finding the relationship between one or more dependent variable and a group of independent variables. In practice, a linear model is usually used for explaining the relationship but other possibilities also exist. Principal Components Regression (PCR), Partial Least Squares Regression (PLSR), Ridge Regression (RR), Stepwise Multiple Linear Regression (SMLR) and Piecewise Linear Regression (PLR) are the most used for linear solutions. For non-linear regression logistic models, growth models, probit and logit models, among others, are used.

PCR and PLR are useful when the matrix does not contain the full model representation. The first step of PCR is the decomposition of the data matrix into latent variables through PCA; then the dependent variable is regressed onto the decomposed independent variables. PLS performs, however, a simultaneous and interdependent PCA decomposition in a way that makes that PLS sometimes handles dependent variables better than does PCR.

Ridge regression (RR) analysis is used when the independent variables are highly interrelated, and stable estimates for the regression coefficients cannot be obtained via ordinary least squares methods (Rozeboom 1979; Pfaffenberger and Dielman 1990). It is a biased estimator that gives estimates with small variance, better precision and accuracy.

The regression procedures need to check some assumptions to their implements. Thus, it is assumed that the residuals are distributed normally. However, it is always a good idea, before drawing final conclusions, to review the distributions of the major variables of interest, in order to inspect the distribution of the residual values (distances of the samples from the estimated model). Furthermore, residuals are also useful for detecting outliers (abnormal data) that is of special interest in authenticity studies.

Outliers are extreme samples that have distorting effects on regression analysis by causing large residual errors or having undue influence on regression coefficients.

Mahalanobis distance, evaluated as \div^2 , is used to discover the outliers among samples and with respect to the solution. Outliers among variables are detected by squared multiple correlation (SMC). In this aspect, multicollinearity (this term describes two variables more or less perfectly correlated and still having similar correlations with the rest of the variables) produces high standard errors on the regression coefficient and estimation is less accurate. Furthermore, there is the singularity. This term applies when some scores are linear combination of others. Thus, multicollinearity and singularity can cause problems in regression analyses, prohibiting or rendering matrix inversion unstable.

Another important question is how many components to use, because using too few components we can build a restricted model (underfitting) whereas using too many the model can be too flexible (overfitting). Cross-validation or Jackknife algorithm (a re-sampling technique based on a "leaving one sample out") can assist to determine the number of components.

The major conceptual limitation of all regression techniques is that one can only ascertain *relationships*, but never be sure about underlying *causal* mechanism. The explanation of conclusions with the assistance of other sciences would avoid reaching nonsense conclusions. A hypothetical paradigm can be to use the electronic nose for detecting the adulteration of refined olive oil with refined seed oils when these kinds of oils do not contain volatiles (refined process of vegetable oils includes the deodorization).

There are other interesting statistical procedures as Canonical Correlation that allows the determination of the correlation between two sets of objects, and to know the explained redundancy between them. Correspondence Analysis, another statistical procedure, is a descriptive/exploratory technique designed to analyse simple two-way and multi-way tables containing some measure of correspondence between the rows and columns. The results provide information that is similar in nature to those produced by factor analysis techniques. Procluster Analysis (Arnold and Williams 1986) is able to analyse the behaviour of each panellist in sensory analysis. Statistical Sensory Wheel (SSW) (Aparicio and Morales 1995), based on directional statistics (Mardia 1972) and PCA, allows clustering inside a circle of not only the sensory descriptors but also the chemical compounds responsible for them. This algorithm is able to analyse the relationship between sensory descriptors and chemical compounds, certain synergies and

antagonisms between chemical compounds, and the interaction between sensory descriptors and basic stimuli (Aparicio *et al.* 1996b).

9.4. Artificial Intelligence methods in food authentication

Statistical packages are generally employed for statistical production purposes because they provide the users with the mechanics of the data analysis but they do not help very much with analytical strategies. Thus, when the analyst faces the authentication problem and see that all the measures have an uncertainty associated with measurement results, he can make the decision of working with methods of artificial intelligence in which the inexactness (generality, ambiguity, vagueness and uncertainty) is the main reason of their existence. The normal question is to wonder which method (neural network, fuzzy logic and expert systems) will give the best conclusions. Each one of these methods (models) has its advantages and disadvantages. Although they are dissected below, simple definitions could light about their possible application in food authentication. Neural networks are models that can learn from past experiences by adaptive programs, which means that they have some of the disadvantages of the discrimination systems. The theory of fuzzy sets provides a convenient means to deal with ill-defined and doubtful data. Expert Systems are self-learning, efficient and reliable programs that are able to solve a particular problem based on knowledge bases of heuristic rules.

9.4.1. Expert Systems

The ability to make decisions on the basis of knowledge distinguishes expert systems (ESs) from statistical programs but their essential characteristics are the self-learning, the look-ahead and the back-propagation up to the point that they are *sine qua non* conditions for true expert systems. Neither statistical procedures nor other artificial intelligence algorithms are able to learn of the past experiences, evaluate decisions before making them (chess's rules) or retry if a decision led the system in a wrong way. The expert systems have to be designed to carry out two main tasks: (i) to extract knowledge rules, and (ii) to design the algorithm controlling them. An ES is composed of a knowledge base and an inference engine (Klahr and Waterman 1986). The former stores the domain knowledge in the form of facts and rules whereas the latter is able to cluster only the

knowledge related to a specific knowledge domain (e.g., a monovarietal virgin olive oil, the geographical origin of a virgin olive oil) and to apply the rules in the appropriate order to infer the correct conclusion. The better the selected information, the greater the experts system success.

ES is based on empirical data and, obviously, the data set should fulfil certain conditions if the process is to be successful; an expert system based on non-refined data would be meaningless. The most important of these conditions is related to the distribution of values over the years and, frequently, to the degree of separability of the foods to be authenticated. The empirical information is used to build some of the knowledge rules.

Expert systems work with rules that have been built independently. Each rule has propositions that are related to a parameter, or a ratio between parameters or an equation classifying categories (e.g., two monovarietal virgin olive oils) of the knowledge domain (e.g., cultivars) to be characterised. The taxonomic organisation of the knowledge takes an arborescent form or tree graph (Aparicio 1988; Zupan *et al* 1988). Each node contains information about the class finding and this information is held in structures called frames. A frame is an abstract specification for a class similar to a property list or record in conventional programming. The frames contain two types of information: domain-specific components and external information. The domain-specific component expresses the substantive characteristics of each parameter while the latter is associated with facts given by the user. Each entity of a frame consists of a name, its attributes and the values linked with these attributes (e.g., *superclass*: aliphatic alcohols; *class*: tetracosanol; *concentration range*: 1.33, 20.37; *geographical origin*: Andalusia). The ES knowledge base is represented by rules obtained from the frame. A rule is represented by means of a set of propositions (premises) that, if they are fulfilled, provides a conclusion associated with a value that indicates its certainty factor (CF): RULE <rule name>, IF <propositions>, THEN <conclusion>; CF= <value>.

There are different types of knowledge rules though the most common are four (Aparicio 1988): inexact reasoning rules, lineal rules, relational rules and heuristic rules. The first type of rules stores the distinctive values of chemical compounds that are characteristics of each cluster of the knowledge domain (e.g. cv. Picual) as premises (e.g., IF the Tridecene concentration is high, THEN the cultivar is Hojiblanca; CF=0.78) The second type of rules stores the coefficients of lineal equations obtained by statistical

procedures working with bivariate distributions. The boundaries of each one of these distributions draw a geometrical figure, so-called 'fuzziness-triangle', where the peak point corresponds to the maximum probability calculated by statistical procedures (e.g., Tridecene concentration (ppm) high [0.612, 0.669, 0.963]; medium [0.342, 0.510, 0.669]; low [0.178, 0.342, 0.510]). The relational rules are quite similar to the previous one although they do not implement an arithmetic calculation among chemical parameters (e.g., IF %Oleic is less than %Linolenic*100, THEN the geographical origin is Andalusia; CF=0.81) (Alonso and Aparicio 1993). The last rule assists the system to determine whether the conclusions pointed out from different frames (e.g, cultivar, geographical origin, extraction systems) are in agreement with one another; for example, if the conclusions are "cv. Koroneiki" (cultivar) and "Greece" (geographical origin), the results would be consistent enough because Koroneiki is a Greek cultivar. If the variety were "cv. Picual" (Spanish cultivar) there would not be any agreement and the system would try to find at other conclusions by following alternative routes in the tree structure (Bundy, 1983; Shirai and Tsujii, 1982)

SEXIA™ is the expert system designed to authenticate virgin olive oil in terms of its geographical origin and cultivars (Aparicio 1988; Aparicio and Alonso 1994; Alonso and Aparicio 1993; Aparicio *et al.* 1994). To date, its knowledge base contains more than 400 reasoning rules that compile around 2000 European virgin olive oil samples characterised by fifty-five chemical compounds. A comparison between the results attained by statistical procedures and this expert system showed that SEXIA™ was 26% better than discriminant analysis in the studies of virgin olive oils produced in Italy (Aparicio *et al.* 1994). Table 1 shows the results of authenticating the geographical origin of virgin olive oil samples produced by the cooperative societies located in the eight regions of the province of Jaén (Spain) that is the highest world producer with an average of about 400,000 tons (1996-1997). Other expert systems have been used or designed by Blaffert (1986), Derde *et al.* (1987), Betteridge *et al.* (1988) and Adler *et al.* (1993).

9.4.2. Neural Networks

The neural networks are essentially non-linear regression models based on a binary threshold unit (McCulloch and Pitts 1943). The structure of neural networks, called a perceptron, consists of a set of nodes at different layers where the node of a layer is linked with all the nodes of the next layer (Rosenblatt 1962). The role of input layer is to feed input

patterns to intermediate layers (also called hidden layers) of units that are followed by an output result layer where the result of computation is read off. Each one of these units is a “neuron” that computes a weighted sum of its inputs from other “neurons” at a previous layer, and outputs a one or a zero according to whether the sum is above or below a certain threshold. The weighted sum of inputs must be or exceed the threshold for unit to fire. Thus, zero means no connection between units whereas positive or negative values mean an excitatory or inhibitory connection between units. This classic definition is generalised in the following equation

$$n_i := g(\sum w_{ij}n_j - \mu_i), \text{ for } j=1 \text{ to } n$$

where n_i represents the state of the unit, $g(x)$ is a general nonlinear function, w_{ij} represents the strength of the connection between unit i and unit j , μ_i is the specific threshold value for unit i , and the sign $:=$ emphasises that the equation is not function of time since it is updated asynchronously by a common clock.

The question is: how does the analyst choose w_{ij} so that a model can carry out a task?. The solution can be the iterative adjustments of the w_{ij} strengths that may be done by supervised or unsupervised learning. The iterative learning procedure based on supervised learning is stopped either after a predetermined number of cycles or after a defined minimum difference between the network output and the correct outputs. For those situations in which information available is extremely reduced, there are particular modelling networks for training stage of the neural networks; for example “learning with a critic” (Barto *et al.* 1991) and “Reward-Penalty” (Barto and Anandan 1985).

The unsupervised procedure does not have a feedback saying how the outputs should be or whether they are correct. This means that unsupervised learning can only do anything useful when there is redundancy in the input data (Barlow 1989). The unsupervised learning can be based on multiple output units that are often active together (e.g., Hebbian learning and extension rules) (Hebb 1949; Oja 1982; Sanger 1989), or on only one output unit per group at a time (e.g., competitive learning). The former learning is very similar to the statistical algorithms of principal components and cluster analysis whereas the competitive learning categorises the input data by models as Willshaw and von der Masburg or Kohonen (Hertz *et al.*, 1991; Kohonen, 1989). The unsupervised learning is rapid as it does not use back-propagation, that can be extremely slow, and it is

advisable to apply it before training a network with a back-propagation supervised procedure.

Another division of neural networks corresponds to the number of layers, a simple perceptron that has only one layer (Minski and Papert 1969) *versus* a multi-layer perceptron that has more than one layer (Hertz *et al.* 1991). This simple differentiation means that the network architecture is very important, and each application requires its own design. To get good results one should store into the network as much knowledge as possible and use criteria for optimal network architecture as the number of units, the number of connections, the learning time, cost and so on. A genetic algorithm can be used to search the possible architectures (Whitley and Hanson 1989).

Table 1 shows the results of comparing the results of authenticating the geographical origin of virgin olive oil samples produced in the regions of Jaén (Spain) by an expert sytem, the supervised procedure of stepwise linear discriminant analysis and the neural network.

Region	No. samples	Correct classifications (%)		
		SEXIA	SLDA	ANN
Campiña	16	99.9	62.5	94.1
Cazorla	8	87.5	25.0	75.0
Condado	8	99.9	75.0	87.5
La loma	17	94.1	82.4	94.1
Martos	19	94.7	84.2	89.5
Sierra Morena	6	99.9	83.3	83.3
Sierra del Segura	12	83.3	58.3	91.7
Sierra Sur	11	90.9	18.2	91.9

Table 9.1. Authentication of the geographical origin of virgin olive oil samples: Comparative results of SEXIA expert system, neural networks and the supervised chemometric procedure of stepwise linear discriminant analysis. Samples collected in the regions of Jaén (Spain)

9.4.3. Fuzzy Logic

In many experimental cases, a certain degree of interference occurs among the measures, which gives rise to possible collections of results but the situation is even more complex if the input data are subjected to uncertainty or imprecision (Kaufman and Gupta 1991). The fuzzy logic is the only mathematical application that can solve properly problems with imprecision in input data.

Fuzzy logic is based on the generalisation of theory of sets 'characteristic function' that Zadeh defined as 'membership function', $\mu(x)$, (Zadeh 1965)

$$\mu_F(x_i) \in [0,1] \quad \forall x_i \in F$$

The function describes the degree of membership, within the closed interval $[0,1]$, to which each element x_i in the knowledge domain belongs to a fuzzy subset, F . The degree of membership 1 means that that x_i definitely does belong to F (full membership), zero means that it does not belong to the fuzzy subset (nonmembership) whereas value between zero and one means partial membership. The relationship between possible values of x_i and their degree of membership can be represented by different graphs as triangular, trapezoidal, bell-shaped or irregular asymmetric functions (Kandel 1986). Based on this function, one can build models of input data into degrees of membership of linguistic fuzzy sets, i.e. low, medium or high; this is called 'fuzzification'. We can further define special modifiers that would act modifying the initial probability distributions; for example, Very X: $\mu_{\text{very } x}(F_i) = \mu^2(F_i)$, and Fairly X: $\mu_{\text{more or less}}(F_i) = \mu^{1/2}(F_i)$. But in order to build a practical system, that must be precise and consistent in its behaviour, one has to impose restrictions with regard to the use of ambiguous modifiers (e.g. very not large) and the order of them in the composition (e.g., very X, fairly X, more or less X, above X, much above X) (Dubois and Prade 1980; Kandel 1986).

However, the identification of the fuzziness associated to single parameter characterizing foodstuffs is not enough. The authentication process is not usually restricted to a single parameter but in fact there are several of them. Whether we want to operate with their membership function (e.g., low linoleic and high 24-methylene cycloarthanol), we need to define operations on the fuzzy set. Thus, the classical rule "R= IF (input) A, THEN (output) B" can be extended to several inputs (antecedents), and perhaps outputs (consequences), linked by the aforementioned logic operations; for example, IF (fairly low linoleic fatty acid) and (very high 24-methylene cycloarthanol), THEN variety is Arbequina (Calvente and Aparicio 1995). The fuzzy relation between pair of sets can also be expressed using basic operations as T-norm, T-conorm, negation, implication and defuzzification (Kandel 1986).

Table 2 shows the results of comparing the results of authenticating three Spanish monovarietal virgin olive oils by the supervised procedure of linear discriminant analysis

and an algorithm of fuzzy logic. The results are promising as the results applying Luckasiewicz T-conorm $S_{1.5}$ (Kandel 1986) are of same order of magnitude than LDA.

Cultivar	No. samples	Correct classifications (%)	
		Fuzzy logic	LDA
Farga	24	91.7	91.7
Hojiblanca	35	62.9	74.3
Picual	228	93.9	96.9

Table 9.2. Authentication of monovarietal virgin olive oils: Comparative results of fuzzy logic algorithms (Calvente and Aparicio 1995) and the supervised chemometric procedure of linear discriminant analysis. Chemical compounds used: linolenic acid, 24-methylenecycloarthanol sterol and copaene hydrocarbon.

REFERENCES

- Adler, B., Schütze, P. and Will, J. (1993). Expert system for interpretation of x-ray diffraction spectra. *Anal Chim Acta* **271**, 287-291.
- Alonso, V. and Aparicio, R. (1993). Characterization of European virgin olive oils using fatty acids. *Grasas Aceites* **44**,18-24.
- Aparicio, R. (1988). Characterization of foods by inexact rules: The SEXIA expert system. *J Chemometr A* **3**, 175-192.
- Aparicio, R. and Alonso, V. (1994). Characterization of Virgin Olive Oils by SEXIA Expert System. *Prog Lipid Res* **33**, 29-38.
- Aparicio, R., Alonso, V. and Morales, M. T. (1994). Detailed and Exhaustive Study of the Authentication of European Virgin Olive Oils by SEXIA Expert System. *Grasas Aceites* **45**, 241-252.
- Aparicio, R., Calvente, J. J. and Morales, M. T. (1996). Sensory Authentication of European Extra-Virgin Olive Oil Varieties by Mathematical Procedures. *J Sci Food Agric* **72**, 435-447.
- Aparicio, R. and Morales, M. T. (1995). Sensory Wheels: A Statistical Technique for Comparing QDA Panels. Application to Virgin Olive Oil. *J Sci Food Agric* **67**, 247-257.
- Aparicio, R., Morales, M. T. and Alonso, V. (1996). Relationship between Volatile Compounds and Sensory Attributes by Statistical Sensory Wheel. *J Am Oil Chem Soc* **73**, 1253-1264.
- Aparicio, R., Morales, M. T. and Alonso, V. (1997). Authentication of European Virgin Olive Oils by Their Chemical Compounds Sensory Attributes and Consumers Attitudes. *J Agric Food Chem* **45**, 1076-1083.
- Aparicio, M.T. Morales, G. Luna, R. Aparicio-Ruiz (2000) Biochemistry and Chemistry of Volatile Compounds Affecting to Consumers' Attitudes of Virgin Olive Oil, in *Flavour and Fragrance Chemistry* (eds. V. Lanzotti and O. Taglialatela-Scafati). Kluwer

Academic Press,; Dordrecht (The Netherlands), 3-14

- Armstrong, R. D. and Beck, P. O. (1990). An algorithm to assist in the Identification of Multiple. Multivariate outliers when using a Least Absolute Value Criterion, in *Robust Regression: Analysis and Applications* (eds. K. D. Lawrence and J. L. Arthur), Marcel Dekker, New York, pp 89-104
- Arnold, G. M. and Williams, A. A. (1986). The use of Generalised Procrustes Techniques in Sensory Analysis, in *Statistical Procedures in Food Research* (ed. J. R. Piggott), Elsevier, London, pp 233-254.
- Baeten, V., *et al.* (1996). Detection of Virgin Olive Oil Adulteration by Fourier Transform Raman Spectroscopy. *J Agric Food Chem* **44**, 2225-2230.
- Barlow, H. B. (1989). Unsupervised Learning. *Neural Comput* **1**, 295-311.
- Barto, A. G. and Anandan (1985). Pattern Recognizing Stochastic Learning Automata. *IEEE Trans Syst Man Cyber* **15**, 360-375.
- Barto, A. G., Sutton, R. S. and Watkins, C. J. C. H. (1991) Learning and Sequential Decision Making, in *Learning and Computational Neuroscience*, (eds. M. Gabriel and J.W. Moore, MIT Press, Cambridge, MA, pp 112-136.
- Betteridge, D., *et al.* (1988). Development of an Expert System for the Selection of Sample Points for Moisture Analysis. *Anal Chem* **60**, 1534-1539.
- Blaffert, T. (1986). EXPERTISE - An expert System for Infrared Spectra Evaluation. *Anal Chim Acta* **191**, 161-168.
- Borg, I. and Lingoes, J. (1987) *Multidimensional similarity structure analysis*. Springer Publishing Co., New York.
- Bundy, A. (1983). *The Computer Modelling of Mathematical Reasoning*, Academic Press, London.
- Calvente, J. J. and Aparicio, R. (1995). A fuzzy filter for removing interferences among membership grade functions. An application to pre-treatment of data in olive oil

authentication. *Anal Chim Acta* **312**, 281-294.

Cooley, W.W. and Lohnes, P.R. (1971). *Multivariate Data Analysis*. John Wiley and Sons, New York.

Derde, M.P., *et al.* (1987). Comparison of Rule-Building Expert Systems with Pattern Recognition for the Classification of Analytical Data. *Anal Chem* **59**, 1868-1871.

Dubois, D. and Prade, H. (1980). *Fuzzy Sets and Systems: Theory and Applications*, Academic Press, San Diego, CA: .

EURACHEM. The fitness for Purpose of Analytical Methods. A Laboratory Guide to method Validation and Related Topics. EURACHEM Secretariat, Teddington UK.

Hartigan, J. A. (1975). *Clustering Algorithms.*, John Wiley and Sons, New York.

Hebb, D.O. (1949). *The Organization of Behaviour*, John Wiley and Sons, New York.

Hertz, J., Krogh, A. and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*, Addison-Wesley Published Co, Reading, MA.

Jacobsen, T. and Gunderson, R.W. (1986) Applied cluster analysis, in *Statistical Procedures in Food Research*, (ed. J.R. Piggot), Elsevier Applied Science, London, pp 361-408.

Kandel, A. (1986). *Fuzzy Mathematical Techniques with Applications*, Addison-Wesley Publishing Co, Reading, MA

Kaufmann, A. and Gupta, M. M. (1991). *Introduction to Fuzzy Arithmetic: Theory and Applications*, van Nostrand Reinhold, New York.

Klahr, P. and Waterman, D. A. (1986). *Expert System Techniques, Tools and Applications*, Addison-Wesley Publishing Co, Reading MA

Kohonen, T. (1989). *Self-Organization and Associative Memory*, Springer-Verlag, Berlin.

Mardia, K. V. (1972). *Statistics of Directional Data*, Academic Press, New York.

- Martens, H. and Næs, T. (1989). *Multivariate Calibration*, John Wiley and Sons, Chichester, UK.
- Massart, D. L. and Kaufman, L. (1983). Hierarchical clustering methods. In *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, (ed. J. D. Winefordner), John Wiley and Sons, New York, pp. 75-99.
- McCulloch, W. S. and Pitts, W. (1943). A logical Calculus of Ideas Immanent in Nervous Activity. *Bull Math Biophys* **5**, 115-133.
- Minski, M. L. and Papert, S. A. (1969). *Perceptrons*, MIT Press, Cambridge, MA.
- Oja, E. (1989). A Simplified Neuron Model as a Principal Component Analyzer. *J Math Biol* **15**, 267-273.
- Pfaffenberger, R. C. and Dielman, T. E. (1990). A Comparison of Regression Estimators when both Multicollinearity and Outliers are present, in *Robust Regression: Analysis and Applications*, (eds. K. D. Lawrence and J.L. Arthur), Marcel Dekker, New York, pp. 243-270.
- Piggott, J. R. and Sharman, K. (1986). Method to aid interpretation of multidimensional data. In *Statistical Procedures in Food Research*, (ed. J.R. Piggott), Elsevier Applied Science, London, pp. 181-132.
- Rosenblatt, F. (1962). *Principles of Neurodynamics*, Spartan, New York.
- Rozeboom, W.W. (1979). Ridge Regression: Bonanza or beguilement?. *Psychol Bull* **86**, 242-249.
- Sanger, T.D. (1989). Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network. *Neural Networks* **2**, 359-473.
- Schiffman, S., Reynolds, M. L. and Young, F. W. (1981). *Introduction to Multidimensional Scaling. Theory, Methods and Applications*, Academic Press, Orlando FL.
- Shirai, Y and Tsujii, J. (1982). *Artificial Intelligence: Concepts, Techniques and Applications*, John Wiley and Sons, Chichester, UK.

Tabachnick, B.G. and Fidell, L.S. (1983). *Using Multivariate Statistics*, Harper and Row, New York

Taylor, B.N., Kuyaat, C.E. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. 1994. National Institute of Standards and Technology, Gaithersburg MD USA.

Wherry, R. J. (1984). *Contributions to Correlational Analysis*, Academic Press, New York.

Whitley, D. and Hanson, T. (1989). Optimizing Neural networks Using Faster More Accurate Genetic Search. Proceeding of Third International Conference on Genetic Algorithms, (eds. J. D. Schaffer. San Mateo CA: Morgan Kaufmann), pp 391-396.

Wold, S. (1978). Cross-validated estimation of the number of components in factor analysis and principal components models. *Technometrics* **20**, 397-406.

Zadeh, L. A. (1965). Fuzzy Sets. *Inf Control* **8**, 338-353.

Zamora, R., Navarro, J. L and Hidalgo, F. J. (1994). Identification and Classification of Olive Oils by High-Resolution ¹³C Nuclear Magnetic Resonance. *J Am Oils Chem Soc* **71**, 361-364.

Zupan, J. (1982). *Clustering of large data sets*, Research Studies Press, New York.

Zupan, J., *et al.* (1988). Building Knowledge into an Expert System. *Chemo Int Lab Syst* **4**, 307-314.

Zupan, J., *et al.* (1994). Classification of multicomponent analytical data of olive oils using different neural networks. *Anal Chim Acta* **292**, 219-234.